W10–W15 *Nucleic Acids Research, 2004, Vol. 32, Web Server issue*
DOI: 10.1093/nar/gkh367

The web server of IBM's Bioinformatics and Pattern Discovery group: 2004 update

Tien Huynh[1] and Isidore Rigoutsos[1,2,*]

[1]Bioinformatics and Pattern Discovery group, IBM T.J. Watson Research Center, PO Box 218, Yorktown Heights, NY 10598, USA and [2]Department of Chemical Engineering, Massachusetts Institute of Technology, Cambridge, MA 02139, USA

Received February 14, 2004; Accepted March 8, 2004

ABSTRACT

In this report, we provide an update on the services and content which are available on the web server of IBM's Bioinformatics and Pattern Discovery group. The server, which is operational around the clock, provides access to a large number of methods that have been developed and published by the group's members. There is an increasing number of problems that these tools can help tackle; these problems range from the discovery of patterns in streams of events and the computation of multiple sequence alignments, to the discovery of genes in nucleic acid sequences, the identification—directly from sequence—of structural deviations from α-helicity and the annotation of amino acid sequences for antimicrobial activity. Additionally, annotations for more than 130 archaeal, bacterial, eukaryotic and viral genomes are now available on-line and can be searched interactively. The tools and code bundles continue to be accessible from http://cbcsrv.watson.ibm.com/Tspd.html whereas the genomics annotations are available at http://cbcsrv.watson.ibm.com/Annotations/.

INTRODUCTION

In the report which appeared in the first installment of Nucleic Acids Research's 'Special Issue on Web Servers', we described in detail several tools that we have web-enabled and made available through the website of the Bioinformatics and Pattern Discovery group at IBM's T. J. Watson Research Center. The tools are implementations of methods that the group's members have designed and published over the course of several years: indeed, since 1996, the focus of the group's research activity has been on the theoretical and applied aspects of pattern discovery with a focus on problems from molecular biology.

In addition to the provision of a web-based graphical user interface to the various tools, executable code bundles for several operating system and processor combinations can be downloaded from the same server and installed locally. Also, we make available content in the form of annotations for more than 130 complete genomes (The figure is accurate as of February 2004.) Provided graphical user interfaces help carry out simple searches of these annotations.

The tools which are currently available are implementations of methods for generic pattern discovery and related applications (1–6), multiple sequence alignment (7,8), gene discovery (9), protein annotation (5,10–13), comparative molecular moment analysis (14,15), the elucidation of non-α-helical conformations in transmembrane helices directly from the amino acid sequence (16), and the *in silico* determination of the potential of peptides to exhibit antimicrobial behavior (Jensen, K., Jung, G., Stephanopoulos, G. and Rigoutsos, I., manuscript submitted).

In light of the discussion found in (17), what follows contains only single-sentence descriptions for those of the services and content which have remained unchanged and, instead, focuses on the new tools and utilities that have been included in the server's latest release.

WHAT IS AVAILABLE ON THE SERVER

In the most recent release of our server, the tools and content are divided into four groups, namely, 'Pattern Discovery Tools (1 of 2)', 'Pattern Discovery Tools (2 of 2)', 'Bio-Dictionary Tools and Content', and 'Other Tools'. These four groups comprise four, four, three and two subchoices respectively—see Figure 1. As in the server's earlier release, each tool's page includes several logical sections: 'parameters', 'options', 'input sequences', 'references', and 'other relevant links', the latter appearing always on the left side of each page. On each and every tool page we also make available context-sensitive help and an email link that permits a user to contact us with questions and comments pertaining to the tools or the operation of the server. In an effort to shorten the learning

*To whom correspondence should be addressed. Tel.: +1 914 945 1384; Fax: 1 914 945 4104; Email: rigoutso@us.ibm.com

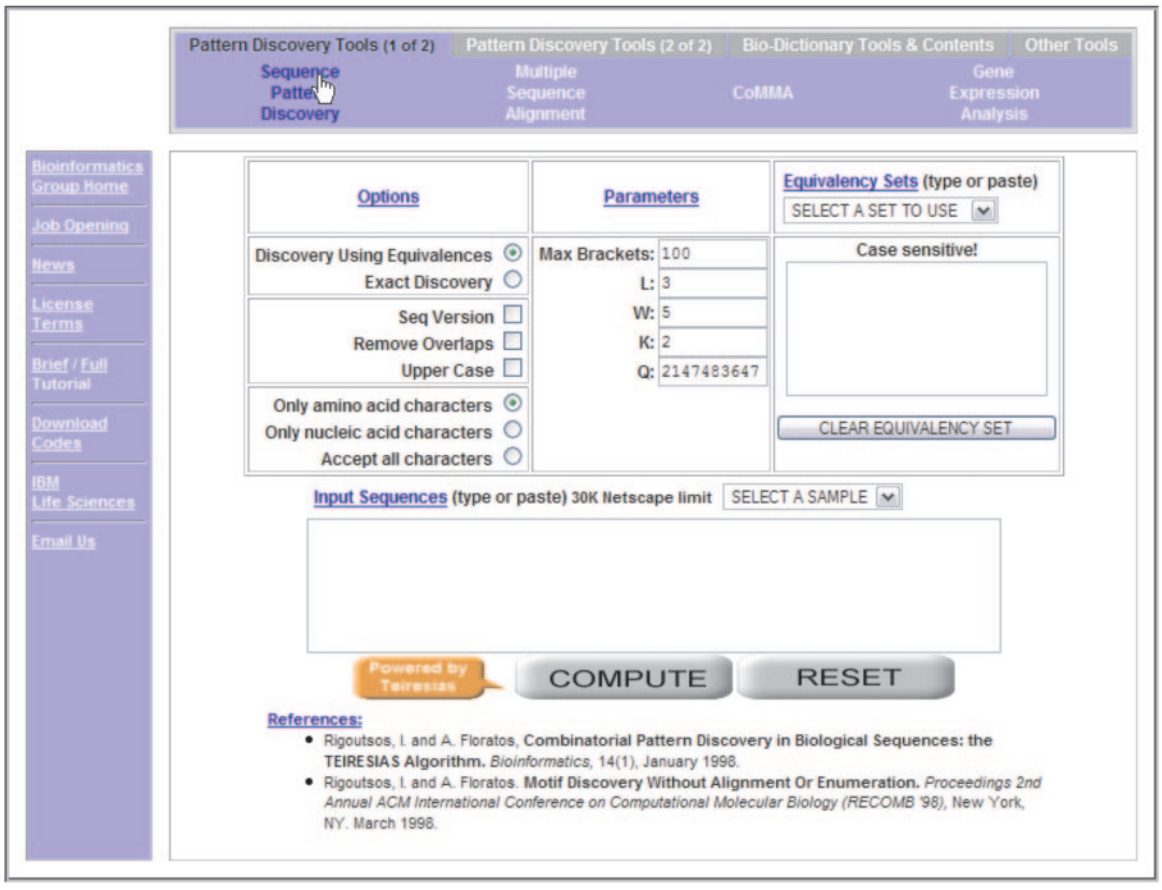

Figure 1. Screen shot of a typical tool page from the most recent release of our web server. The collection of available tools is now divided among four different groups: 'Pattern Discovery (1 of 2)', 'Pattern Discovery (2 of 2)', 'Bio-Dictionary Tools and Content' and 'Other Tools'. See text for more details.

phase for the user, we continue to provide sample input datasets accessible through pull-down menus, while the options and parameters are automatically initialized to default values. Additionally, real-time checking of the various settings and choices by the user helps us to identify logical and other conflicts and subsequently prompt with appropriate suggestions.

Executable codes for all our tools and for a variety of operating systems and CPU combinations are available through http://cbcsrv.watson.ibm.com/download.phtml.html. The latest code bundles include the complete graphical user interface of our web server so that the user can replicate most of the server's functionality on his or her local machine. The only two tools that continue to require remote access are the protein annotation and gene discovery ones; this limitation is due to the very large sizes of the underlying data tables that the tools use. However, the data tables are made available to users upon request.

For some of the provided tools, the command-line codes contained in last year's (2003) release required the execution of several utilities in a proper order. Realizing that this is a cumbersome manner in which to run the tools, we have augmented our code distribution this year with the inclusion of single-command versions of the web-based tools: this in turn permits a user to make the tools part of an existing pipeline, obviating the need to interact with a graphical user interface.

Both the downloadable code bundles and the web-based services are provided free of charge for those users carrying out not-for-profit work (e.g. academic or personal research). Use of the tools in a for-profit, commercial environment requires a license, and additional details on the existing terms and conditions can be found on our web site.

We next describe each of the tools that is included in the server's 2004 release. As already mentioned, only minimal information is provided for those tools that have already been described in (17). On the other hand, the newly added tools and new functionality are described in considerable detail.

Sequence Pattern Discovery (http://cbcsrv.watson.ibm.com/Tspd.html). This is our implementation of Teiresias, a two-phase, combinatorial algorithm for general-purpose pattern discovery (1–3). The tool can be used to process amino acid sequences, nucleic acid sequences, ASCII representations of numerical data, and so on.

Multiple Sequence Alignment (http://cbcsrv.watson.ibm.com/Tmsa.html). This tool represents our implementation of the MUSCA algorithm, which allows the user to align multiple streams of letters in order to reveal any salient features that may be present in the considered input (7,8).

Gene Expression Analysis (*Time Series*) (http://cbcsrv.watson.ibm.com/Tgea.html). This tool is used to analyze datasets that track the induction/repression of genes over time and in response to exogenous changes (5).

Gene Expression Analysis (*Association Data*)/*Association Discovery* (http://cbcsrv.watson.ibm.com/Tad.html). This tool

is used to process general inputs representing association data, and to report patterns that are guaranteed to be maximal in composition and length (5). See also below for a new addition to the tool's graphical user interface.

Comparative Molecular Moment Analysis (*CoMMA*) (http://cbcsrv.watson.ibm.com/Tco.html). CoMMA is used to process information from moments of molecular mass and moment expansions of electrostatic potentials up through and inclusive of second order and to derive molecular descriptions that remain invariant under Euclidean transformations (14,15). For multiple molecule inputs, Teiresias can be optionally run on the CoMMA descriptors to determine commonalities among subsets of the input molecules.

Integer Pattern Discovery (http://cbcsrv.watson.ibm.com/Tipd.html). This tool is the integer-based implementation of Teiresias and is to be used with event streams of positive integers.

Natural Text Mining—Word units (http://cbcsrv.watson.ibm.com/Ttwpd.html). This tool uses the integer version of Teiresias to process natural text (not necessarily English) and treats each input word as a unit of information; the provided input should be in ASCII, and can appear in free format.

Natural Text Mining—Symbol units (http://cbcsrv.watson.ibm.com/Ttspd.html). The only difference between the previous and this tool is that here the unit of information is the individual character.

Protein Annotation (http://cbcsrv.watson.ibm.com/Tpa.html). This tool allows the user to automatically annotate a given amino acid sequence using the Bio-Dictionary-based approach that we described in (11,13).

Gene Identification (http://cbcsrv.watson.ibm.com/Tgi.html). This tool provides access to the implementation of our dictionary-based approach to finding genes in prokaryotic genomes (9); its graphical user interface permits the linking of the tool's output to the protein annotation tool at the click of a button.

Human Cytomegalovirus (HHV5) Annotation (http://cbcsrv.watson.ibm.com/virus/). This site is an entry point to a database that we have compiled and which summarizes our work on annotating those ORFs of the virus which are presumed to code for a protein; the two efforts whose results are included in this database are discussed in (18,19).

Genome Annotations (http://cbcsrv.watson.ibm.com/Annotations/). Using a recent installment of the Swiss-Prot/TrEMBL database (20) with ~875 000 sequences, we recomputed an updated version of the Bio-Dictionary (11), which now comprises approximately 57 million patterns. The new Bio-Dictionary was subsequently employed to generate *updated* annotations for the 70 complete genomes that we reported in (17) and also to annotate an *additional* 60 new genomes from the archaeal, bacterial and eukaryotic domains. Annotations for genomes to be released at a later point in time will be added to our website as they become available.

Non-canonical Conformations of Transmembrane Helices (http://cbcsrv.watson.ibm.com/Ttkw.html). This tool represents a recent addition to our tool collection and permits one to locate and characterize non-canonical conformations of transmembrane helices directly from sequence. Transmembrane helices of polytopic proteins are common building elements of biologically important structures such as tissue- and/or ligand-specific receptors and enzymes. In earlier work (21), we reported that non-canonical conformations occur frequently in these proteins and are critical in determining the proteins' structure and function. We subsequently sought to develop a method that would permit the discovery and characterization of such deviations from α-helicity as belonging to one of three types: tight turn, wide turn or true kink. The method which we devised is pattern-based and exhibits great sensitivity and specificity; details and a discussion can be found in (16). When presented with a sequence in FASTA format, the tool determines the locations and extent of all identifiable instances of tight/wide turns and true kinks and reports the results in the form of a plot that shows the support of each instance as a function of position. Figure 2 shows the output generated for a C-terminal fragment of the bovine

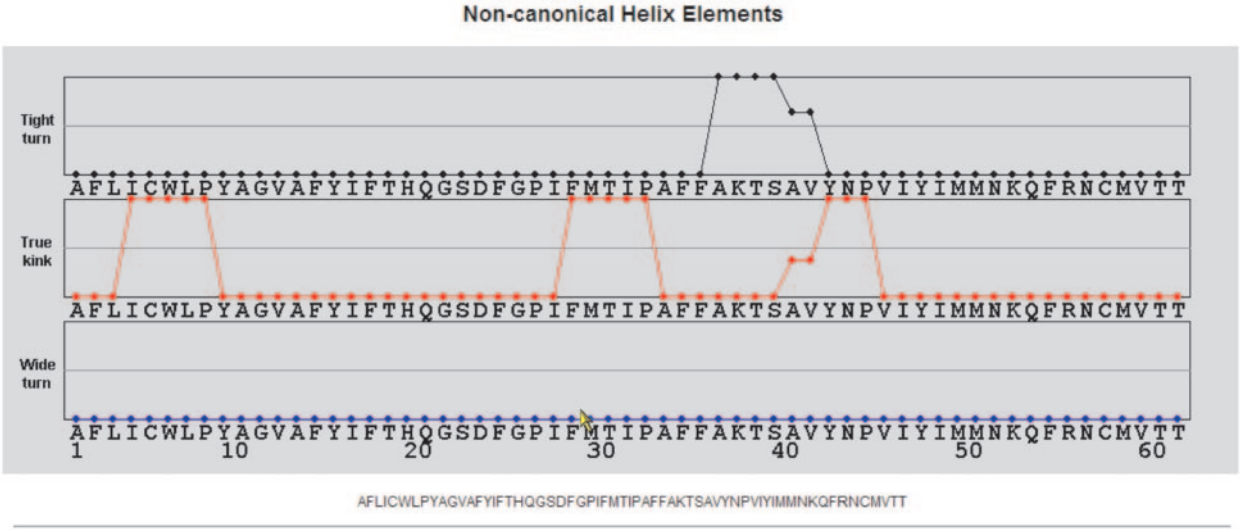

Figure 2. The output page of the tool which localizes and characterizes non-canonical conformations of transmembrane helices, directly from sequence. The input here is a C-terminal fragment of the bovine rhodopsin OPSD_BOVIN. The computed results are in agreement with the deviations that can be observed in the reported crystal structure of this protein. See also text for more information.

rhodopsin (OPSD_BOVIN). The *i*-th position of the x-axis corresponds to the *i*-th amino acid of the input sequence, whereas the y-axis indicates the degree of confidence in the reported instance. Notice the interweaving of the three non-canonical conformations between positions 27 and 45 of the input sequence.

Antimicrobial Protein Annotation (http://cbcsrv.watson.ibm.com/Tamp.html). In recent years, the proliferation of bacterial strains which are resistant to multiple antibiotics such as penicillin, tetracyclins and vancomycin has led to a growing interest in the study of antimicrobial peptides (22). Current estimates put the cost of treating and preventing infections by these strains to several billion dollars in the United States alone; but clearly in an era of fast transportation by means of air travel the existence of such pathogens represents a global concern. Antimicrobial peptides are short amino acid sequences that are ubiquitous among multicellular organisms and have a net positive charge and an amphipathic 3-dimensional structure. The latter two properties result in their affinity for the negatively charged outer membranes of bacteria, which in turn leads to an increase in the membranes' permeability and the eventual lysis of the cell. Because of the great potential that antimicrobial peptides represent, we have been focusing on the development of *in silico* methods for studying these peptides. This tool implements a method that we developed recently (Jensen, K., Jung, G., Stephanopoulos, G. and Rigoutsos, I., manuscript submitted) and which, when presented with an amino acid sequence, can determine whether any parts of it have the potential to exhibit antimicrobial activity. Moreover, the method can also determine which of the known antimicrobial peptide families is likely to be the closest to the query sequence in terms of antimicrobial behavior. Figure 3 shows only part of the output page that this new tool will generate when presented with the sequence of DEFN_HUMAN, the human neutrophil defensin. The reported results include a plot that indicates antimicrobial behavior as a function of position within the query sequence, local alignments to known antimicrobial peptides, direct linking to Swiss-Prot/TrEMBL's Expasy web server, which in turn permits the retrieval of complete Swiss-Prot/TrEMBL records for sequences of interest, an enumeration of those patterns from the collection used by our search engine that connect

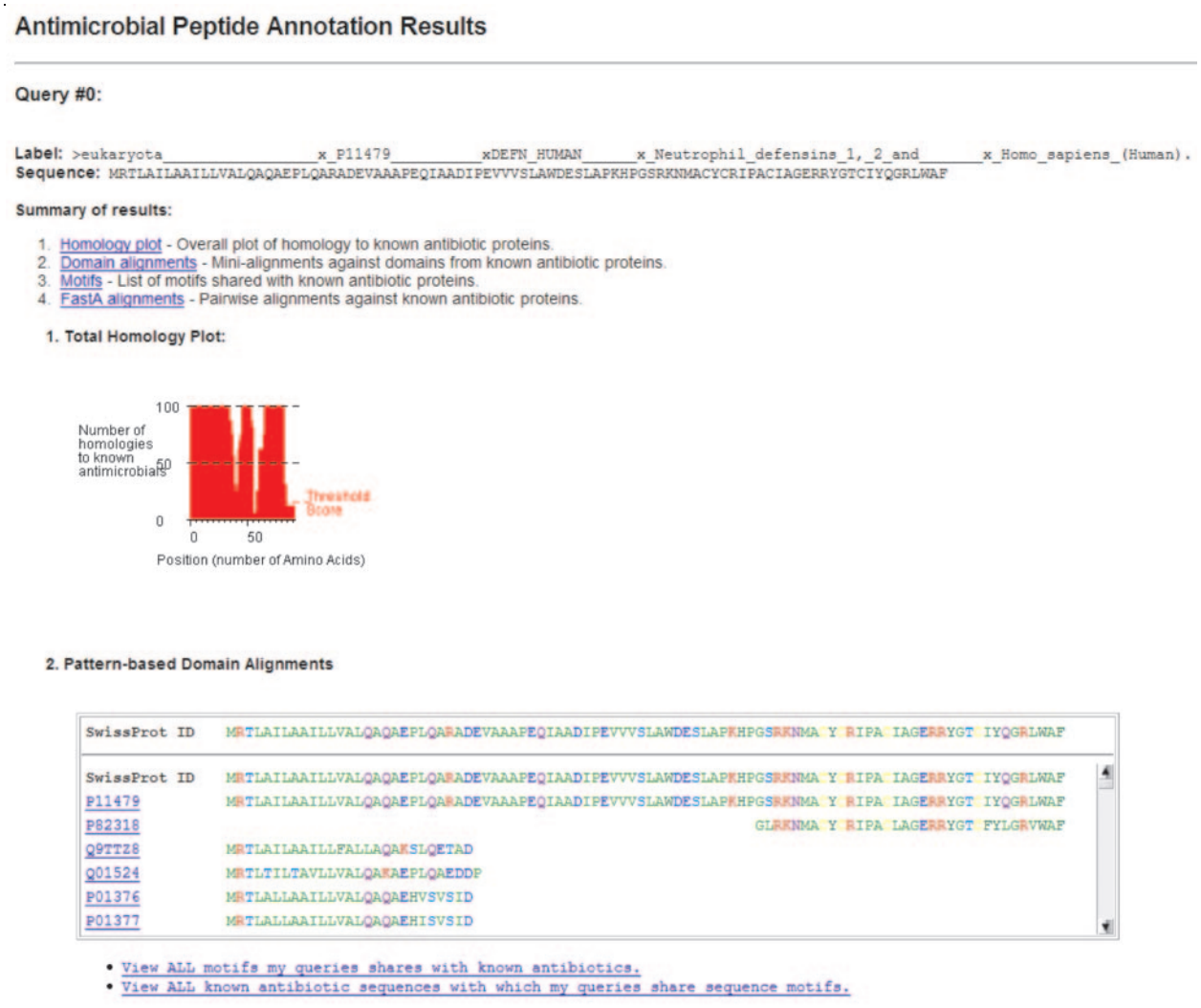

Figure 3. Partial output of the newly developed tool for studying antimicrobial peptides. The query in this case is DEFN_HUMAN, a neutrophil defensin from *Homo sapiens*. The actual output page includes much more information than is shown here. See also text for a discussion.

parts of the query sequence to known antimicrobial peptides and, finally, the results of searching the query sequence in our library of known antimicrobial using Fasta (23).

Other Novel Components—enhanced interface for one of the tools. In addition to the two new tools that we describe above, we have also expanded the graphical user interface and capabilities of the Gene Expression Analysis (Association Data)/Association Discovery tool. In particular, the user has now the option to provide labeled inputs where the rows and columns of the array of numbers to be processed have alphanumeric labels (of up to 20 characters). When the user selects one of the discovered patterns for plotting, we return (i) a plot showing the subsets of rows and columns that form the pattern and (ii) the string corresponding to the first row that participates in the pattern, with the numerical entries indicating the pattern clearly delineated. Additionally, a provided button that is labeled 'summarize' permits the user to interactively display only the submatrix of rows and columns corresponding to the pattern; notably, the elements of the submatrix will be appropriately labeled if the original input contains labels for the rows and columns—see Figure 4.

Other Novel Components—parallel Teiresias. Given the ever increasing size of the public databases with biological information, we have devoted a lot of our attention to generating a parallel version of the Teiresias algorithm, since the latter underlies many of the methods that we have developed and make available through our web site. In fact, we have completed an implementation of the algorithm that is suitable for both shared memory and message passing architectures (Huynh, T., Parida, L., Platt, D. and Rigoutsos, I., manuscript submitted). We currently make this implementation available as part of the downloadable code bundles. A matching graphical user interface that will permit users to run sequence pattern discovery tasks on multiple processors is currently in preparation and will be added to our web server prior to the publication of this article.

PLANNED EXTENSIONS AND UPCOMING TOOLS

We conclude by briefly outlining three more of our ongoing efforts toward further enhancing the usefulness of this server. The first is a tool that we are in the process of preparing for web-enablement and which implements the algorithm described in (24). This algorithm allows the discovery of so-called π-patterns, a special type of patterns whose elements can be permuted across the patterns' instances. For example, ABCDE and ACBED are instances of the same π-pattern

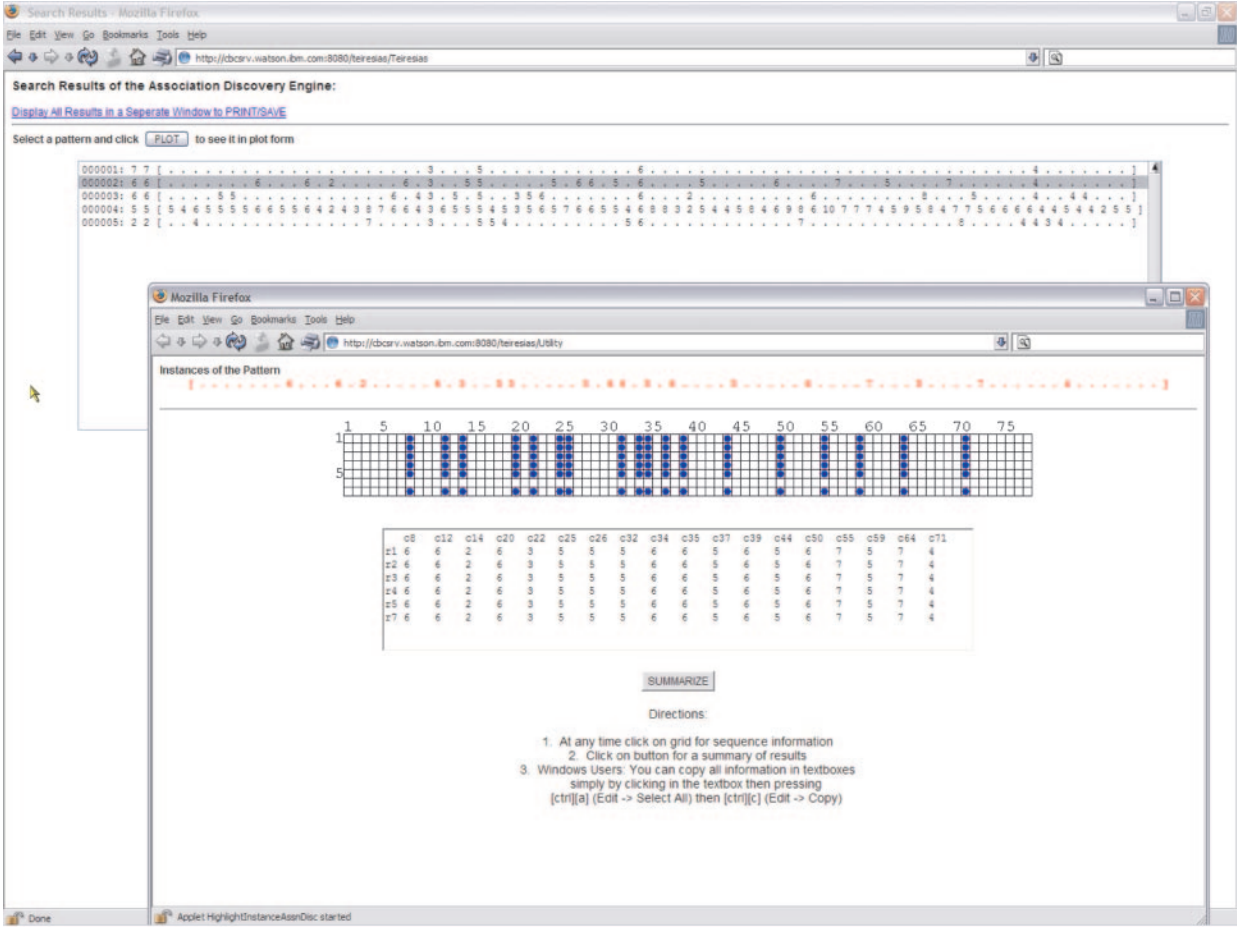

Figure 4. Sample output of the Gene Expression Analysis (Association Data)/Association Discovery tool. An newly developed and enhanced graphical user interface permits the speedy evaluation of the results and the determination of the actual values which participate in the discovered patterns. Row and column labels, optionally provided by the user, can be used to indicate the identity of each pattern's components. See text for additional information.

which comprises A, B, C, D and E. This algorithm is particularly useful for comparative genomics analysis (in this case, a π-pattern would be composed of multiple genomic regions), in comparing proteins (here a π-pattern would comprise 'domains'), and elsewhere.

The second effort is a tool that we are developing and which can deduce the secondary structure of a protein directly from its amino acid sequence. The key idea has its underpinnings in the method we have developed for protein annotation (13) and revolves around the use of patterns with secondary structure attributes that have been derived from available crystallographic data.

Finally, the third effort whose fruits will be appearing on our web server relates to our working toward augmenting the repository of full-genome annotations with searching capabilities that exploit the power of the IBM DB2 database management system: not only will search speeds improve, but also we will provide users with the ability to issue complicated queries across the entire repository of on-line protein annotations currently corresponding to 130 genomes.

CONCLUSION

We have provided an update of the capabilities of the IBM Bioinformatics group's web server as it stands at the beginning of 2004. Since its first description in (17), we have augmented the server by making available more tools and almost doubling the genomic content that is made available through it. Currently, a total of 12 tools are remotely accessible through a web-based graphical user interface. Also, we make available recently computed annotations for nearly twice as many full genomes as we reported in (17): more than 130 archaeal, bacterial, eukaryotic and viral genomes are currently available on-line.

ACKNOWLEDGEMENTS

The authors thank Mashama McFarlane for his contributions to the development of various aspects of the graphical user interfaces.

REFERENCES

1. Rigoutsos,I. and Floratos,A. (1998) Combinatorial pattern discovery in biological sequences: the *Teiresias* algorithm. *Bioinformatics*, **14**, 55–67.
2. Rigoutsos,I. and Floratos,A. (1998) Motif discovery without alignment or enumeration. *Proceedings of the Second Annual ACM International Conference on Computational Molecular Biology (RECOMB)*, New York, NY, pp. 221–227.
3. Floratos,A. and Rigoutsos,I. (1998) On the time complexity of the TEIRESIAS algorithm. *IBM Technical Report, RC 21161 (94582)*. IBM TJ Watson Research Center, New York.
4. Parida,L., Rigoutsos,I., Floratos,A, Platt,D.E. and Gao,Y. (2000) Pattern discovery on character sets and real valued data: linear bound on irredundant motifs and an efficient polynomial time algorithm. *Proceedings of the 11th Annual ACM/SIAM Symposium on Discrete Algorithms (SODA '00)*. San Francisco, CA, pp. 297–308.
5. Rigoutsos,I., Floratos,A., Parida,L. P., Gao,Y. and Platt,D. (2000) The emergence of pattern discovery techniques in computational biology. *Metab. Eng.*, **2**, 159–177.
6. Parida,L., Rigoutsos,I. and Platt,D.E. (2001) An output-sensitive flexible pattern discovery algorithm. *Proceedings of the 12th Annual Symposium on Combinatorial Pattern Matching*, Jerusalem, Israel, pp. 131–142.
7. Parida, L., Floratos,A. and Rigoutsos,I. (1998) MUSCA: an algorithm for constrained alignment of multiple data sequences. *Proceedings of the Ninth Workshop on Genome Informatics*, Tokyo, Japan, pp. 112–119.
8. Parida,L., Floratos,A. and Rigoutsos,I. (1999) An approximation algorithm for alignment of multiple sequences using motif discovery. *J. Comb. Optim.*, **3**, 247–275.
9. Shibuya,T. and Rigoutsos,I. (2002) Dictionary-driven microbial gene finding. *Nucleic Acids Res.*, **30**, 2710–2725.
10. Floratos, A., Rigoutsos,I., Parida,L., Stolovitzky,G. and Gao,Y. (1999) Sequence homology detection through large-scale pattern discovery. *Proceedings of the Third Annual ACM International Conference on Computational Molecular Biology (RECOMB '99)*, Lyon, France, pp. 164–173.
11. Rigoutsos,I., Floratos,A., Ouzounis,C., Gao,Y. and Parida,L. (1999) Dictionary building via unsupervised hierarchical motif discovery in the sequence space of natural proteins. *Prot. Struct. Funct. Genet.*, **37**, 264–277.
12. Rigoutsos,I., Gao,Y., Floratos,A. and Parida,L. (1999) Building dictionaries of 1D and 3D motifs by mining the unaligned 1D sequences of 17 archaeal and bacterial genomes. *Proceedings of the Seventh International Conference on Intelligent Systems for Molecular Biology (ISMB '99)*, Menlo Park, California. AAAI Press, pp. 223–233.
13. Rigoutsos,I., Huynh,T., Floratos,A., Parida,L. and Platt,D. (2002) Dictionary-driven protein annotation. *Nucleic Acids Res.*, **30**, 3901–3916.
14. Silverman,B.D and Platt,D.E. (1995) Comparative molecular moment analysis (CoMMA): 3D Qsar without molecular superposition. *J. Med. Chem.*, **39**, 2129.
15. Platt,D.E. and Silverman,B.D. (1996) Registration, orientation, and similarity of molecular electrostatic potentials through multipole matching. *J. Comput. Chem.*, **17**, 358.
16. Rigoutsos,I., Riek,P., Graham,R.M. and Novotny,J. (2003) Structural details (kinks and non-α conformations) in transmembrane helices are intrahelically determined and can be predicted by sequence pattern descriptors. *Nucleic Acids Res.*, **31**, 4625–4631.
17. Huynh,T., Rigoutsos,I. Platt,D., Parida,L. and Shibuya,T. (2003) The web server of IBM's bioinformatics and pattern discovery group. *Nucleic Acids Res.*, **31**, 3645–3650.
18. Novotny,J., Rigoutsos,I., Coleman,D. and Shenk,T. (2001) *In silico* structural and functional analysis of the human cytomegalovirus (HHV5) genome. *J. Mol. Biol.*, **310**, 1151–1166.
19. Rigoutsos,I., Novotny,J., Huynh,T. Chin-Bow,S. Parida,L., Platt,D., Coleman,D. and Shenk,T. (2003) *In silico* pattern-based analysis of the human cytomegalovirus (HHV5) genome. *J. Virol.*, **77**, 4326–4344.
20. Boeckmann,B., Bairoch,A., Apweiler,R., Blatter,M.-C., Estreicher,A., Gasteiger,E., Martin,M.J., Michoud,K., O'Donovan,C., Phan,I., Pilbout,S. and Schneider,M. (2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.*, **31**, 365–370.
21. Riek,R.P., Rigoutsos,I., Novotny,J. and Graham,R.M. (2001) Non-a-helical elements modulate polytopic membrane architecture. *J. Mol. Biol.*, **306**, 349–362.
22. Zasloff,M. (2002) Antimicrobial peptides of multicellular organisms. *Nature*, **415**, 389–395.
23. Pearson,W.R. (1996) Effective protein sequence comparison. *Meth. Enzymol.*, **266**, 227–258.
24. Eres,R., Landau,G. and Parida,L. (2003) A combinatorial approach to automatic discovery of cluster patterns. *Proceedings of the Workshop on Algorithms for Bioinformatics*. Budapest, Hungary.